An Optimization-Based Approach to Patient Grouping for Acute Healthcare in Australia

A.M. Bagirov[1] and L. Churilov[2]

[1] Centre for Informatics and Applied Optimization, School of Information Technology and Mathematical Sciences, The University of Ballarat, Victoria 3353, Australia
a.bagirov@ballarat.edu.au
[2] School of Business Systems, Monash University, Victoria, 3800, Australia
leonid.churilov@infotech.monash.edu.au

Abstract. The problem of cluster analysis is formulated as a problem of nonsmooth, nonconvex optimization, and an algorithm for solving the cluster analysis problem based on the nonsmooth optimization techniques is developed. The issues of applying this algorithm to large data sets are discussed and a feature selection procedure is demonstrated. The algorithm is then applied to a hospital data set to generate new knowledge about different patterns of patients resource consumption.

1 Introduction

The subject of cluster analysis is the unsupervised classification of data and discovery of relationships within the data set without any guidance. The basic principle of identifying these hidden relationships are that if input patterns are similar, then they should be grouped together. Two inputs are regarded as similar if the distance between these two inputs (in multidimensional input space) is small. Due to its highly combinatorial nature, clustering is a technically difficult problem. Different approaches to the problem of clustering analysis that are mainly based on statistical, neural network and machine learning techniques have been suggested in [5,6,8,9,12]. An excellent up-to-date survey of existing approaches is provided in [4].

Mangasarian ([7]) describes an approach to clustering analysis based on the bilinear programming techniques. Bagirov et al ([2]) propose the *global optimization* approach to clustering and demonstrate how the supervised data classification problem can be solved via clustering. The objective function in this problem is both nonsmooth and nonconvex and this function has a large number of local minimizers. Problems of this type are quite challenging for general-purpose global optimization techniques. Due to a large number of variables and the complexity of the objective function, general-purpose global optimization techniques, as a rule, fail to solve such problems.

It is very important, therefore, to develop optimization algorithms that allow the decision maker to find "deep" local minimizers of the objective function. Such "deep" local minimizers provide a good enough description of the dataset

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2659, pp. 20–29, 2003.

under consideration as far as clustering is concerned. The optimization algorithm discussed in this paper belongs to this type and is based on nonsmooth optimization techniques. The proposed approach has two distinct important and useful features:

- it allows the decision maker to successfully tackle the complexity of large datasets as it aims to reduce both the number of data instances (records) and the number of data attributes (so-called "feature selection") in the dataset under consideration without loss of valuable information
- it provides the capability of calculating clusters step-by-step, gradually increasing the number of data clusters until termination conditions are met.

The power of this approach is illustrated by conducting the clustering analysis of a hospital data set for the purposes of generating new knowledge about patient resource consumption. Knowledge about resource consumption and utilization is vital in modern healthcare environments. In order to manage both human and material resources efficiently, a typical approach is to group the patients based on common characteristics. The most widely used approach is driven by the Case Mix funding formula, namely to classify patients according to diagnostic related groups (DRGs). Although it is clinically meaningful, some experience suggests that DRG groupings do not necessarily present a sound basis for relevant knowledge generation ([10,11]).

The *objective* of this paper is, therefore, two-fold:

- to describe the new approach to clustering analysis based on nonsmooth nonconvex optimization techniques
- to demonstrate how optimization-based clustering techniques can be utilized to suggest an alternative grouping of the patients that generates homogeneous patient groups with similar resource utilization profiles.

Demographics, admission, and discharge characteristics are used to generate the clusters that reveal interesting differences in resource utilization patterns. Knowledge that is not available from DRG information alone can be generated using this clustering method since demographic and other data is used for the patient grouping. This knowledge can then be used for prediction of resource consumption by patients. A detailed case study is presented to demonstrate the quality of knowledge generated by this process. It is suggested that the proposed approach can, therefore, be seen as an evidence-based predictive tool with high-knowledge generation capabilities.

The rest of the paper is organized as follows: nonsmooth optimization approach to clustering is presented in Sect. 2; Sect. 3 describes an algorithm for solving clustering problem; the issues of the complexity reduction for clustering in large data set are discussed in Sect. 4; the description of the emergency dataset is given in Sect. 5, while Sect. 6 presents the discussion of the numerical experiments; Sect. 7 concludes the paper.

2 The Nonsmooth Optimization Approach to Clustering

In this section we present a formulation to the clustering problem in terms of nonsmooth, nonconvex optimization.

Consider set A that consists of r n-dimensional vectors $a^i = (a_1^i, \ldots, a_n^i),\ i = 1, \ldots, r$. The aim of clustering is to represent this set as the union of q clusters. Since each cluster can be described by a point that can be considered as the center of this cluster, it is instrumental to locate a cluster's center in order to adequately describe the cluster itself. Thus, we would like to find q points that serve as centers of corresponding clusters.

Consider now an arbitrary set X, consisting of q points $x^1, \ldots, x^q$. The distance $d(a^i, X)$ from a point $a^i \in A$ to this set is defined by

$$d(a^i, X) = \min_{s=1,\ldots,q} \|x^s - a^i\|$$

where

$$\|x\|_p = \left(\sum_{l=1}^{n} |x_l|^p \right)^{1/p}, \ \ 1 \le p < +\infty, \qquad \|x\|_\infty = \max_{l=1,\ldots,n} |x_l|.$$

The deviation $d(A, X)$ from the set A to the set X can be calculated using the formula

$$d(A, X) = \sum_{i=1}^{r} d(a^i, X) = \sum_{i=1}^{r} \min_{s=1,\ldots,q} \|x^s - a^i\|.$$

Thus, as far as optimization approach is concerned, the cluster analysis problem can be reduced to the following problem of mathematical programming

$$\text{minimize} \ \ f(x^1, \ldots, x^q) \quad \text{subject to } (x^1, \ldots, x^q) \in \mathbb{R}^{n \times q}, \tag{1}$$

where

$$f(x^1, \ldots, x^q) = \sum_{i=1}^{r} \min_{s=1,\ldots,q} \|x^s - a^i\|. \tag{2}$$

If $q > 1$, the objective function (2) in the problem (1) is nonconvex and nonsmooth. Note that the number of variables in the optimization problem (1) is $q \times n$. If the number q of clusters and the number n of attributes are large, the decision maker is facing a large-scale global optimization problem. Moreover, the form of the objective function in this problem is complex enough not to become amenable to the direct application of general purpose global optimization methods. Therefore, in order to ensure the practicality of the optimization approach to clustering, the proper identification and use of local optimization methods with the special choice of a starting point is very important. Clearly, such an approach does not guarantee the globally optimal solution the problem (1). On the other hand, this approach allows one to find a "deep" minimum of the objective function that, in turn, provides a good enough clustering description of the dataset under consideration.

Note also that the meaningful choice of the number of clusters is very important for clustering analysis. It is difficult to define *a priori* how many clusters represent the set A under consideration. In order to increase the knowledge generating capacity of the resulting clusters, the optimization based approach discussed in this paper adopts the following strategy: starting from a small enough number of clusters q, the decision maker has to gradually increase the number of clusters for the analysis until certain termination criteria motivated by the underlying decision making situation is satisfied. Further discussion on this issue with specific suggestions on the number of clusters is given in Sect. 4.

From optimization perspective this means that if the solution of the corresponding optimization problem (1) is not satisfactory, the decision maker needs to consider the problem (1) with $q + 1$ clusters and so on. This implies that one needs to solve repeatedly arising global optimization problems (1) with different values of q - the task even more challenging than solving a single global optimization problem. In order to avoid this difficulty, a step-by-step calculation of clusters is implemented in the optimization algorithm discussed in the next section.

Finally, the form of the objective function in the problem (1) allows one to significantly reduce the number of records in a dataset. The way the proposed algorithm utilizes this feature is discussed in more detail in Sects. 4 and 6.

3 An Optimization Algorithm for Solving Clustering Problem

In this section we describe an algorithm for solving cluster analysis problem in a given dataset.

Algorithm 1. *An algorithm for solving cluster analysis problem.*

Step 1. (Initialization). Select a tolerance $\epsilon > 0$. Select a starting point $x^0 = (x_1^0, \ldots, x_n^0) \in IR^n$ and solve the minimization problem (5). Let $x^{1*} \in IR^n$ be a solution to this problem and f^1 be the corresponding objective function value. Set $k = 1$...

Step 2. (Computation of the next cluster). Select a point $x^0 \in IR^n$, construct a new starting point $x^{02} = (x^{1*}, x^0) \in IR^{2n}$, and solve the following minimization problem:

$$\textit{minimize} \quad f^k(x) \quad \textit{subject to} \quad x \in IR^n \tag{3}$$

where

$$f^k(x) = \sum_{i=1}^{m} \min\{\|x^{1*} - a^i\|, \ldots, \|x^{k*} - a^i\|, \|x - a^i\|\}.$$

Step 3. Let $x^{k+1,*}$ *be a solution to the problem (3). Take*

$$x^{k0} = (x^{1*}, \ldots, x^{k*}, x^{k+1,*})$$

as a new starting point and solve the following minimization problem:

$$\text{minimize} \quad f^k(x) \quad \text{subject to} \quad x \in IR^{(k+1)\times n} \tag{4}$$

where

$$f^k(x) = \sum_{i=1}^{m} \min_{j=1,\ldots,k+1} \|x - a^i\|.$$

Step 4. (Stopping criterion). Let $x^{k+1,*}$ *be a solution to the problem (4) and* $f^{k+1,*}$ *be the corresponding value of the objective function. If*

$$\frac{f^{k,*} - f^{k+1,*}}{f^{1*}} < \epsilon$$

then stop, otherwise set $k = k + 1$ *and go to Step 2.*

Both problems (3) and (4) are nonsmooth optimization problems and we use the discrete gradient method developed in [1] to solve them.

4 Complexity Reduction for Large-Scale Data Sets

As was mentioned earlier in this paper, due to the highly combinatorial nature of clustering problems, two characteristics of a given data set can severely affect the performance of a clustering tool: a number of the data records (instances) and a number of data attributes (features). In simple terms, if each record in the data set is a separate row, the former is equal to the number of "rows", while the latter is equal to the number of "columns" of data in the data set. The natural priorities of a decision maker in this case are to reduce both the number of features and the number of instances without loss of knowledge generating ability.

In order to reduce the number of features, the feature selection procedure discussed below can be implemented.

4.1 Feature Selection

If in (1) $q = 1$, (3) becomes the following convex programming problem:

$$\text{minimize} \quad f(x) = \sum_{i=1}^{r} \|x - a^i\| \quad \text{subject to } x \in \mathbb{R}^n. \tag{5}$$

The solution x^* to the problem (5) is then the center of the set A. Definitely, one center cannot give good enough description of the set A from clustering perspective; however, it contains some information about the structure of this set.

On the other hand, the problem (5) is a classical convex programming problem and there exist effective and fast methods for its solution. Bagirov et al ([3]) discusses a feature selection algorithm for supervised data classification. This algorithm calculates the centers of each class in a dataset under consideration by solving problem (5) and removes closest coordinates in a step-by-step fashion as long as the structure of classes remains unchanged.

In the case of an unsupervised data classification, the feature selection is more difficult to achieve. One of the possible approaches to the feature selection for unsupervised classification is to use at least one of the features as an outcome and the rest of the features as inputs. In this case one can apply the feature selection algorithm from [3] to achieve the desired outcome.

4.2 Number of Instances Reduction

The form of the objective function in the problem (1) allows one to significantly reduce the number of vectors $a^i,\ i = 1, \ldots, m$. Let D be a $m \times m$ matrix (i, j)-th element d_{ij} where $d_{ij} = \|a^i - a^j\|$. This is a symmetric matrix. For each $i,\ i = 1, \ldots, m$, calculate

$$r_i = \min_{j \neq i} d_{ij}.$$

and

$$r_0 = \sum_{i=1}^{m} r_i.$$

Select $\epsilon = cr_0$ so that $c > 0$ is some number. Empirical results of numerical experiments show that the best values for c are $c \in [0, 2]$. Then, the following simple procedure to reduce the number of vectors $a^i,\ i = 1, \ldots, m$ can be used: select first vector a^1, remove from the data set all the vectors for which $d_{1j} \leq \epsilon$, and assign to this vector a number of removed vectors. Then select the next remaining vector and repeat the above procedure for this vector, *etc.* Results of numerical experiments reported below suggest that such a procedure allows one to significantly reduce the number of instances in the data set.

5 Case Study

The data for this study was obtained from Frankston Hospital, a medium sized health unit located in the South-Eastern suburbs of Melbourne, Australia. It is part of a network of hospitals which serves nearly 290,000 people year-round, with a seasonal influx of visitors to the area of up to 100,000. The area is a prime seaside retirement location where there is a high proportion of elderly people. Demand for the services of Frankston hospital is exacerbated during holiday periods, when visitors impact heavily on emergency services.

The data used in this study consists of 15,194 records of admitted patients for years 1999/2000. It contains information about demographics of patients and their length of stay (LOS) in the hospital.

Age, gender and initial admissions information are known when the patients arrive at the hospital. Patients can be admitted to various wards including emergency, children's ward, cardiac care ward, short stay unit, post natal unit, hospital-in-the-home unit, intensive care unit, etc. Time spent in emergency data are obtained during the patient stay. Information about patients' DRGs, LOS and last discharge wards are determined once the patient leaves the hospital. There are about one thousand different DRG groups that are coded using the 3 digit ICD-9 (International Classification of Disease) codes. LOS of a patient is calculated as the difference between the time the patient is admitted to the hospital and the time the patient is discharged from the hospital.

As mentioned in earlier discussion, numerical experiments were carried out to determine the optimal number of clusters that could adequately distinguish the data. During the experiments LOS was not used as an input in order to determine whether each cluster would exhibit different LOS characteristics.

The analysis of the underlying decision making situation suggests the following criteria for selecting the "optimal" number of clusters:

- *The clusters themselves are distinct in terms of LOS* (if two groups of patients could independently come up with two different average LOS, then chances are these two groups are different from one another)
- *The variables that belong to each cluster make sense* (the variables that each cluster has should be distinct and carry some information of its own. When each cluster is analyzed, its profile should be unique and meaningful).
- *The sizes of clusters are comparable* (the size of each cluster needs to be monitored. If the cluster is too large then it is possible that more distinct groups could lie in the cluster. Likewise if it is too small, then there is high probability that the cluster is artificial).

Once the clusters are obtained, each cluster profile is examined in detail, and an examination of the DRGs with each cluster can then be performed.

6 Results and Discussion

As discussed earlier in the paper, in order to use the feature selection algorithm suggested in [3], one of the features has to be used as an output. LOS is the best candidate for such a job because it not only allows one to get a good enough partition in the datasets, but also was used in the similar capacity in [10].

The dataset is divided into separate classes using different values of LOS and then the feature selection algorithm suggested in [3] is used to reduce the length of the data records. Despite the fact that several different values of LOS have been used to divide the dataset, the subset of most informative features remains the same for all cases. This subset includes the following five features: time the patient spends in emergency department, patient's age, patient's gender, the "admitted to" ward, and the "discharged from" ward. Taking into consideration the fact that the initial statistics included 13 different related parameters, the feature selection procedure reduces the number of features more than twice,

thus significantly reducing the complexity of the corresponding optimization problems.

During the next stage of investigation, the LOS and patient's diagnostic group (DRG) are added to this list and a new dataset is formed. Algorithm 1 is then applied to the latter dataset to identify clusters. We use Euclidean norm in our experiments. The tolerance is set to $\epsilon = 10^{-2}$ and $c = 1.5$ in numerical experiments. Four features including LOS, patient's time in emergency department, age, and gender are selected as inputs at this stage and then the distribution of patients with respect to DRG and hospital wards is analyzed for individual clusters. Since gender is a categorical feature and has only two values, the dataset can be divided into two subsets containing only males/females for which clustering is performed, and then the results for these two subsets are compared.

The "male" subset of the data set contains 7659 instances and the "female" contains 7509 instances. The number of instances is reduced using the procedure discussed in Sect. 4. Selecting $c = 1.5$ enabled this procedure to reduce the number of "male" and "female" instances to 2766 and 1363 instances respectively. Note that the total number of instances in the reduced data set is 4129, meaning that by applying the number of instances reduction procedure, the size of the data set is reduced by more than 3 times.

Then the clustering algorithm is applied to calculate clusters in both subsets. In both cases the algorithm identified 8 clusters. Further reduction in the tolerance parameter ϵ led to the appearance of two new very small and insignificant clusters. According to the principles for the number of clusters selection outlined in the previous section, the conclusion is that these data sets contain only 8 meaningful clusters. Different values of c can be used in numerical experiments. For $c \in [0, 1.5)$, the structure of clusters is very similar to the one generated for $c = 1.5$. On the other hand, for $c > 2$ quite a different structure is observed that does not satisfy the criteria for clusters specified in the previous section.

For the "male" subset the first cluster contains elderly patients (average age is 73) whose LOS is relatively small (average value is 13 hours) and who are admitted and discharged from the emergency department. The second cluster contains elderly patients (average age is 72) whose LOS is very large (average value is 697 hours). The third cluster contains young patients (average age is 26) with a relatively small time in the emergency department (average value is 0.8 hours). The fourth cluster contains young patients (average age is 28) whose time in the emergency department is average (average value is 3.9 hours). Note that in both cases the admitting and discharging wards are almost the same. The fifth cluster contains elderly patients (average age is 73) whose LOS (average value is 294 hours) and time spent in emergency department is large enough (average value is 5.5 hours) and who use a nonhomogeneous mix of admitting and discharging wards. The sixth cluster contains elderly patients (average age is 71) whose LOS is large enough (average value is 250 hours) whereas time in emergency department is very small (average value is 0.9 hours) and who also use a nonhomogeneous mix of admitting and discharging wards. The seventh cluster

contains more middle-aged patients (average age is 59) whose LOS is not large (average value is 89 hours) while time in emergency department is large (average value is 7.0 hours), who are typically admitted to and discharged from the special care wards. Finally, the last cluster contains middle aged patients (average age is 61) whose LOS and time in emergency department are not large (average values are 121 and 2.1 hours respectively) and who use a nonhomogeneous mix of admitting and discharging wards.

A very similar situation is observed as the result of the analysis of "female" subset of the dataset.

It is very important to note that despite the fact that different clusters contain very different kinds of patients as far as their resource consumption is concerned, the distribution of DRGs for different clusters is very similar. This observation strongly suggests that DRG alone is not capable of adequately differentiating the patients based on their resource consumption and therefore should not be used as a single basis for hospitals funding.

7 Conclusions

In this paper a nonsmooth nonconvex optimization-based algorithm for solving cluster analysis problem has been proposed. As this algorithm calculates clusters step by step, it allows the decision maker to easily vary the number of clusters according to the criteria suggested by the nature of the decision making situation not incurring the obvious costs of the increased complexity of the solution procedure. The suggested approach utilizes the stopping criterion that prevents the appearance of small and artificial clusters. The form of the objective function allows one to significantly reduce the number of instances in a dataset - the feature that is extremely important for clustering in large scale data sets.

The power of this approach has been illustrated by conducting the clustering analysis of a hospital data set containing over 15,000 records for the purposes of generating new knowledge about patient resource consumption. Knowledge that is not available from DRG information alone has been generated using this clustering method. This knowledge can be used by hospital managers for prediction of resource consumption by different patients. The proposed approach can, therefore, be seen as an evidence-based predictive tool with high-knowledge generation capabilities.

References

[1] Bagirov, A.M.: Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices, In: Eberhard, A., Hill, R., Ralph, D. and Glover, B.M. (eds.): Progress in Optimization: Contribution from Australasia, Kluwer Academic Publishers (1999) 147–175

[2] Bagirov, A.M., Rubinov, A.M. and Yearwood, J.: Using global optimization to improve classification for medical diagnosis and prognosis. Topics in Health Information Management **22** (2001) 65–74

[3] Bagirov, A.M, Rubinov, A.M. and Yearwood, J.: A heuristic algorithm for feature selection based on optimization techniques. In: Sarker, R., Abbas, H. and Newton, C.S. (eds.): Heuristic and Optimization for Knowledge Discovery, Idea Publishing Group, Hershey (2002) 13–26
[4] Jain, A.K., Murty, M.N. and Flynn, P.J.: Data clustering: a review. ACM Computing Surveys **31(3)** (1999) 264–323
[5] Dubes, R. and Jain, A.K.: Clustering techniques: the user's dilemma. Pattern Recognition **8** (1976) 247–260
[6] Hawkins, D.M., Muller, M.W. and ten Krooden, J.A.: Cluster analysis, In: Hawkins, D.M.: Topics in Applied Multivariate Analysis, New York, Cambridge University press (1982)
[7] Mangasarian, O.L.: Mathematical programming in data mining. Data Mining and Knowledge Discovery **1** (1997) 183–201
[8] McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. New York, John Wiley (1992)
[9] Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (eds.): Machine Learning, Neural and Statistical Classification. Ellis Horwood Series in Artificial Intelligence, London (1994)
[10] Ridley, S. Jones, S., Shahani, A., Brampton, W., Nielsen, M. and Rowan, K.: Classification Trees. A possible method for iso-resource grouping in intensive care. Anaesthesia **53** (1998) 833–840
[11] Siew, E.-G., Smith, K. and Churilov, L.: A neural clustering approach to iso-resource grouping for acute healthcare in Australia. Proceedings of the 35th International Conference on Systems and Systemics, Hawaii (2002)
[12] Spath, H.: Cluster Analysis Algorithms. Ellis Horwood Limited, Chichester (1980)